HHS Public Access

Author manuscript

Br J Math Stat Psychol. Author manuscript; available in PMC 2019 November 14.

Published in final edited form as:

Br J Math Stat Psychol. 2019 November ; 72(3): 466–485. doi:10.1111/bmsp.12169.

Using Multidimensional IRT to Evaluate How Response Styles Impact Measurement

Daniel J. Adams[1], **Daniel M. Bolt**[1], **Sien Deng**[2], **Stevens S. Smith**[3], **Timothy B. Baker**[3]

[1]Department of Educational Psychology, University of Wisconsin, Madison

[2]ACT, Inc.

[3]Department of Medicine, University of Wisconsin, Madison

Abstract

Multidimensional item response theory (MIRT) models for response style (e.g., Bolt, Lu, & Kim, 2014; Falk & Cai, 2016) provide flexibility in accommodating various response styles, but often present difficulty in isolating the effects of response style(s) from the intended substantive trait(s). In the presence of such measurement limitations, we consider several ways in which MIRT models are nevertheless useful in lending insight into how response styles may interfere with measurement for a given test instrument. Such a study can also inform whether alternative design considerations (e.g., anchoring vignettes, self-report items of heterogeneous content) that seek to control for response style effects may be helpful. We illustrate several aspects of an MIRT approach using real and simulated analyses.

Various item-response theory (IRT)-based approaches have been proposed for the study of response styles in self-report rating-scale instruments. A recent paper made a distinction between multi-process tree IRT and threshold-based IRT models (Böckenholt & Meiser, 2017). We consider this a useful distinction not only in highlighting model types, but also in contrasting different motivations that may underlie response-style modeling. Multi-process tree methods (Böckenholt, 2012; DeBoeck & Partchev, 2012; Jeon & DeBoeck, 2016; Plieninger & Meiser, 2014; Thissen-Roe & Thissen, 2013) typically provide a more prescriptive characterization of how respondents select among rating-scale categories on an item that, as a result, is seemingly associated with a clearer separation of the influence of substantive and response-style traits (whether valid or not). Consider for example a response pattern in which the highest possible score category on a Likert-type rating scale (e.g., “strongly agree”) was selected for all items (and where agreement is assumed to imply a higher level on the substantive trait). A multi-process tree model likely concludes that the respondent is at a high level on both the substantive trait (due to consistent agreement with all items) and an extreme-response-style (ERS) trait (due to consistent subsequent selection of the extreme category across all items). By contrast, threshold models (e.g., Bolt & Johnson, 2009; Eid & Rauber, 2000; Falk & Cai, 2016; Wetzel & Carstensen, 2015) are generally agnostic to response process. Selection among response categories is often portrayed as being simultaneously influenced by both substantive and response-style traits

Corresponding author: Daniel Bolt, 1025 W. Johnson, Room 859, Madison WI USA (dmbolt@wisc.edu).

and, as a consequence, observed item scores are generally less informative indicators of each trait type. Under a threshold approach, respondents may arrive at the exact same item score pattern for different reasons. Given our hypothetical response pattern containing only "strongly agree" responses, under a threshold model the pattern may reflect a respondent with a high level on both substantive and ERS traits (i.e., a low threshold for extreme response), as in the multi-process tree model. But the pattern may also reflect a respondent with a moderate level on the substantive trait and high ERS, or a respondent with a high substantive-trait level and no ERS (i.e., a moderate threshold for an extreme response). As a result, when used to estimate respondent scores, a threshold approach might be expected to yield a higher degree of uncertainty (e.g., higher standard errors) in both the estimated substantive and response-style traits.

In their review of response styles and psychometric methods, Baumgartner and Steenkamp (2001) considered many potential theoretical causes for response styles, an issue that is likely important to consider in selecting among methods. One possibility is that a response style reflects a respondent's attempt to reduce the cognitive demand associated with distinguishing several categories of a rating scale. For example, a respondent presented with seven rating categories for each item may primarily use only a subset of the available scores (e.g., 1, 4, and 7) not as the result of a deliberate process, but as a means of simplifying the task of distinguishing multiple levels of agreement on the rating scale. Such a theory coincides with speculation that response styles are more apparent on longer survey instruments (Kieruj & Moors, 2012). One consequence of this theory is the potential for numerous different response-style types. Indeed, studies that have taken an exploratory approach to response-style modeling have in fact often identified a large number of detectable response-style types (Bolt, Lu, & Kim, 2014; Van Rosmalen, Van Herk, & Groenen, 2010).

Such complexities may make response style modeling a greater challenge under a multiprocess-tree approach. In this paper we seek to illustrate how a threshold approach can be conducive to an exploratory study of response styles and response-style effects. To this end, we consider a model that Bolt, Lu and Kim (2014) originally presented in the context of an anchoring vignette design that we demonstrate can also be applied without vignettes. In addition to the flexibility afforded in modeling several response-style types, we suggest that a potentially underutilized application of threshold-response-style models relates to ways that the models can inform both how, and to what extent, response styles interfere with measurement. Such interference in the measurement of a substantive trait may arise due to the uncertainty response styles create above and beyond their potential contributions to bias. The goal of this paper is to examine several applications for which an exploratory threshold approach to response-style modeling may be well-suited, including: (1) an understanding of response styles as sources of IRT person misfit; (2) an evaluation of the implications of response-style heterogeneity on reduced precision in measurement of the substantive trait; and (3) an ability to compare different test designs (e.g., tests with parallel versus heterogeneous items) in terms of the consequences of particular forms of response-style heterogeneity, as well as response-style detection and control.

The threshold approach taken in this paper makes minimal assumptions regarding the underlying cause(s) of response style heterogeneity, effectively allowing any form of response heterogeneity to emerge for individual respondents, while also allowing for more explicit quantifications of how response style contribute to uncertainty in measurement. One advantage of quantifying the uncertainty is that it can also help in quantifying the potential value of adopting an alternative design that seeks to measure and control for response styles. Although such alternatives are not studied in this paper, they include, for example, (1) anchoring vignettes (King, Murray, Salomon, & Tandon, 2004), namely objective scenarios that respondents rate using the same rating scale as used for the self-report items, or (2) self-report items of heterogeneous content (Greenleaf, 1992) to measure and potentially control for response-style heterogeneity. The identification of consequential response-style heterogeneity may also encourage consideration of a different response format, such as forced-choice or situational-judgment format types. In this respect, psychometric models used to inform about the effects of response style are possibly useful even if not used as scoring models.

A Multidimensional Item Response Model for Response Style

Bolt, Lu, and Kim (2014) considered a multidimensional nominal response model (MNRM) for response style in which the probability of selecting response category k (=1, …, l) is modeled using continuous latent variables to represent both substantive and response-style traits. The Bolt et al. (2014) model accounts for response style using a respondent-level category vector, denoted $\boldsymbol{\eta}_j = (\eta_{j1}, \eta_{j2}, \ldots \eta_{jl})$, that represents a respondent's disproportionate propensities toward use of each of the rating scale categories. The probability of respondent j selecting category k on item i is given by:

$$P(X_{ji} = \mathrm{k} \mid \theta_j; \eta_j) = \frac{\exp(a_{ik}\theta_j + c_{ik} + \eta_{jk})}{\Sigma_{h=1}^{l} \exp(a_{ih}\theta_j + c_{ih} + \eta_{jh})} \quad (1)$$

where arbitrary normalization constraints $\Sigma_{h=1}^{l} c_{ih} = 0$ and $\Sigma_{h=1}^{l} \eta_{jh} = 0$ are applied within item and respondent, respectively. The values of $\boldsymbol{\eta}_j$ reflect a respondent's disproportionate tendency to over- or under-select individual categories, conditional upon the unidimensional substantive trait (θ_j). A positive η_{jk} suggests the respondent over-selects category k while a negative η_{jk} implies under-selection. For the application we present below, we considered l = 7, and fixed the category slopes (i.e., the scoring function) for each item i as $\boldsymbol{a}_i$ = (−3,−2,−1,0,1,2,3). We chose this scoring function to produce a latent metric for the substantive trait that has a similar interpretation to the traditional sum score metric. These fixed slopes are commonly used as a measurement constraint (e.g., Masters, 1982). The model in (1) was applied in Bolt, Lu and Kim (2014) in the context of an anchoring vignette design; for the purpose of this paper, however, we considered an application of the model without vignettes. Our study of the model without anchoring vignettes by simulation suggested that for a suitably long test (>50 items), the variability in respondent-level η_{jk} estimates appears to be captured sufficiently well for the types of applications to be considered in this paper (recovery correlation between true and estimated η_{jk} >.75). Although not a focus of this paper, additional studies of the model in (1) with and without

the application of vignettes would appear useful, both in better understanding the model, optimal vignette designs (e.g., how many vignettes appear necessary), and the implications of traditional assumptions needed when using vignette designs (von Davier, Shin, Khorramdel, & Stankov, 2017).

Similar to Bolt, Lu and Kim (2014), in this paper we estimated the model using a Markov chain Monte Carlo (MCMC) algorithm. We adopted priors $c_{ik} \sim Normal(0,1)$ for the unnormalized item category parameters; for the respondent parameters we used $\theta_j \sim Normal(0,1)$ for the substantive trait, and for the unnormalized η parameters we assumed $\eta_{jk} \sim Normal(0, \sigma^2_\eta)$, where $\sigma^2_\eta \sim \text{Gamma}(1, 1)$. We chose to independently sample the parameters across categories recognizing the alternative possibility of multivariate priors. One reason for adopting this approach was to help ensure an equivalence in the detection of response styles of any form, an issue of particular relevance to the person-fit analyses. As the variance of the θ prior was fixed at 1 to identify the metric, we estimated the variance of the η parameters (σ^2_η). As we show shortly, this quantity proves useful as an index for quantifying the degree of response-style heterogeneity in the population. We discuss this index further in later sections.

We implemented the MCMC algorithm using WinBUGS 1.4 (Spiegelhalter, Thomas, & Best, 2003). Specification of the priors above leads to a Metropolis Hastings sampling algorithm that used an initial 4000 iterations to tune proposal distributions. We subsequently simulated up to 6000 additional iterations for purposes of estimating model parameters. Convergence of the chain to a posterior distribution was monitored using the Gelman-Rubin (1992) R^2 criterion for the item intercept and respondent parameter estimates. We reported as point estimates of respondent parameters $(\hat{\theta}, \hat{\eta})$ the mean of the observations sampled from the final 5000 iterations (omitting the first 1000 of the 6000 estimation iterations as burn-in), and used posterior standard deviations (psd's) as indicators of precision.

Study 1: Response Styles as Sources of IRT Person Misfit

The MNRM in (1) can potentially be used to study response styles at the individual respondent level. We applied the MNRM here in conjunction with a person-fit evaluation to examine the nature and the extent of response-style heterogeneity observed in the 68-item Wisconsin Inventory of Smoking Dependence Motives (WISDM-68; Piper, Felderman, Piasecki, Bolt, Smith, Fiore, & Baker, 2004).

Study 1 Goals

One concern regarding response-style heterogeneity is its potential to undermine the validity of response patterns as indicative of the substantive trait. In item response theory, one means of investigating validity at the respondent level uses person fit-analyses that, within a Bayesian framework, can be examined using deviance statistics (Glas & Meijer, 2003). In this study, we define relevant deviance statistics that permit evaluations of over-/under-use of rating scale categories.

Study 1 Methods

Posterior predictive model checking (PPMC) entails a comparison of deviance statistics as defined for the observed data $T(x_j)$ against replicated data sets $T(x_{rep,j})$ that conform to the model being fit (Sinharay, 2005). To investigate misfit related to response-style heterogeneity of the type captured by the model in (1), we defined deviance statistics with respect to a partial credit model (PCM; Masters, 1982):

$$P(X_{ji} = \mathrm{k} \mid \theta_j) = \frac{\exp(a_{ik}\theta_j + c_{ik})}{\Sigma_{g=1}^{l} \exp(a_{ig}\theta_j + c_{ig})} \tag{2}$$

where $\boldsymbol{a}_i = (-3,-2,-1,0,1,2,3)$. The PCM is viewed as a special case of the model in (1) where $\boldsymbol{\eta}_j = \mathbf{0}$ for all respondents. Although our misfit analysis is based on application of the PCM, to further explore the nature of the response-style heterogeneity, we examine these deviance-statistic results in relation to the MNRM estimates observed for individual respondents so as to understand the occurrence of person misfit. Similar to how we estimate the model in (1), for the PCM in (2) we adopted priors of $\theta_j \sim Normal(0,1)$ across respondents and $c_{ik} \sim Normal(0,1)$ across items and categories, where $\Sigma_{h=1}^{l} c_{ih} = 0$, and replicated the same MCMC estimation process. At each iteration, we also generated replicate response vectors $x_{rep,j}$ for each respondent j, and a corresponding deviance statistic $T(x_{rep,j})$. Across iterations we used $T(x_{rep,j})$ to define respondent-level reference distributions against which the deviance statistic for the observed response pattern $T(x_j)$ was compared.

The first deviance statistic we considered was the log-likelihood of the response pattern for respondent j:

$$T_L(x_j) = \sum_{i=1}^{NI} \sum_{k=1}^{l} x_{ijk} \log P_{ijk}$$

where $P_{ijk} = P(X_{ji} = \mathrm{k} \mid \theta_j)$ and $x_{jik} = 1$ if $X_{ji} = \mathrm{k}$ and 0 otherwise. A second deviance statistic aligning more with the form of response-style heterogeneity in (1) considers the sum of the squared deviations between the observed (OF) and expected (EF) category frequencies at the respondent level. OF is an l-element vector representing the number of times each category was selected (across items) by the respondent. EF is a vector of the same dimension indicating the expected frequency the respondent was to select each category conditional upon the respondent θ under the PCM. The EFs served as a reference against which we evaluated the OFs for the observed response vectors. Specifically, the deviance statistic for respondent j was calculated as:

$$T_F(x_j) = \sum_{k=1}^{l} (OF_{j,k} - EF_{j,k})^2$$

To calculate the T_F statistic, we first needed to determine $EF_{j,k}$ for each respondent and score category. For this purpose, we applied an initial MCMC simulation of 1000 iterations

under the PCM model with the sole purpose of estimating $EF_{j,k}$ for each respondent and category. The expected frequency of category k for respondent j was then the sum of probabilities of the respondent selecting category k across the NI items:

$$EF_{j,k} = \sum_{i=1}^{NI} P_{ijk}.$$

As $EF_{j,k}$ was calculated at each iteration of the chain, we took the mean across the last 5000 iterations as a point estimate against which the observed and replicate frequencies were calculated. Subsequently, a second simulated chain of 10,000 iterations was run in which the observed frequencies, based on the simulated replicate response vectors, were calculated in relation to the estimated $EF_{j,k}$ from the previous run.

Model fit was determined by the proportion of 10,000 replications for which $T(x_j) > T(x_{rep,j})$ under the T_L statistic and for which $T(x_j) < T(x_{rep,j})$ under the T_F statistic. We were interested not only in respondent under-fit but also over-fit, as overly consistent response patterns involving repeated selection of the same category(ies) may yield response patterns with higher than expected conformity to the PCM model. As a result, we flagged respondents for which either the proportion is less than or equal to 0.025 (under-fitting) or greater than or equal to 0.975 (over-fitting).

For purposes of illustration, we applied these methods to actual response data from a 68-item measure of tobacco dependence, the Wisconsin Inventory of Smoking Dependence Motives (WISDM-68; Piper et al., 2004). The WISDM-68 items reflect statements indicative of tobacco dependence (e.g., “I often smoke without thinking about it”, “I frequently crave cigarettes”) that each respondent scores on a 7-point scale (1 = “not at all true of me” to 7 = “extremely true of me”). An overall score is reported as either a mean or sum score across all items. Higher scores indicate higher levels of dependence. The current dataset involves the item responses of 1504 smokers administered the WISDM-68 prior to a quit attempt in a tobacco cessation study. A small number of missing responses, < .5%, were recoded as 4s for purposes of the current analysis. We discuss this issue further below.

Study 1 Results

Table 1 displays cross-tabulation results for each of the 1,504 observed response patterns in the WISDM-68 data. As seen from the row and column marginals in the table, substantial differences appear to exist in the sensitivities of the T_L and T_F statistics to person misfit. Specifically, a much larger proportion of respondents showed under-fit with respect to T_F compared to T_L. These findings were not surprising in the presence of actual response-style heterogeneity, as the T_F statistic was designed to be much more sensitive to such effects than T_L. As (in practice) greater attention tends to be devoted to under-fitting patterns when attending to likelihood-based criteria, we note that well over half of the response patterns (821/1504) displayed under-fit with respect to T_F, while less than 10% (140/1504) of the same patterns showed under-fit for T_L. A total of 92/140 (66%) of patterns showing under-fit under T_L also showed under-fit under T_F, a higher conditional proportion than those either

fitting or over-fitting under $T_L(\frac{729}{1364} = 53\%; \chi^2(1, N = 1504) = 7.69, p < .01)$, suggesting response style is a frequent source of likelihood-related misfit where it occurs. At the same time, many of the under-fitting patterns under T_F were not detected as either under-fitting or over-fitting under T_L, suggesting that over-/under-utilization of response categories often does not yield likelihood-based misfit. The difference between deviance statistics underscores the fact that response-style heterogeneity will often not manifest as likelihood-related misfit at the respondent level.

Table 2 provides examples of observed response patterns along with their corresponding PPMC p-values for the T_L and T_F statistics, and parameter estimates under the MNRM. The first three patterns provide examples of patterns that were detected as under-fitting for T_F, but not for T_L. By attending to the profile of η estimates observed under the model in (1), the nature of the response style can be interpreted. Respondent 1472 produced a pattern reflective of ERS, with large positive estimates for η_1 and η_7, while Respondent 1228 showed a tendency to avoid the extreme categories, to overselect categories 3 and 4(with largest positive estimate for η_4) and to underselect categories 2 and to a lesser extent 6. Other patterns were more similar to those of Respondents 720, 230, and 1431 and displayed patterns not reflective of any of the more traditionally discussed response style types. For example, Respondent 720 displayed over-selection of 2s and 6s (as reflected by large positive estimates of η_2 and η_6); Respondent 230 showed over-selection of 2s and under-selection of 5s (as reflected by large positive estimate of η_2 and a large negative estimate of η_5), while Respondent 1431 involved over-selection of 3s (as reflected by large positive estimate of η_3). Importantly, the standard errors for the η estimates in all of these patterns were relatively large implying the specific degree of response style is difficult to measure precisely at the category level; nevertheless, general patterns of over- and under-utilization of response categories does seem to be discernible for many of the patterns.

Further exploration of response-style heterogeneity can attend to the covariance structure of $\eta_j = (\eta_{j1}, \eta_{j2}, \ldots, \eta_{jI})$ across respondents observed under the MNRM. In this analysis, we extracted principal components of the covariance matrix of the $\boldsymbol{\eta}$ estimates after fitting the MNRM model.[1] We observed rescaled eigenvalues of 3.05, 1.37, 1.13, .72, .69, .04 and 0 (the last being 0 due to the sum-to-zero constraint applied to η), indicating approximately 79 percent of the variance is explained by 3 components. The component loadings of the first three components are shown in Table 3. It is important to note that the components are best viewed as continuous dimensions of response style, with both ends of the component continua reflecting different (and diametrically opposing) aspects of response style. With this in mind, Component 1 appeared to correspond to an ERS/anti-ERS dimension, while Component 2 distinguished respondents over-selecting Category 2 and under-selecting Categories 4 and 5 (at the positive end of the component) versus the opposite (at the negative end). Finally Component 3 distinguished respondents that over-selected Category 6 and under-selected 3 and 4 (at the positive end of the component) versus the opposite (at the

[1]Principal components could also be extracted from a latent $\boldsymbol{\eta}$ covariance structure if such a structure were introduced as a part of the model. When we specify a model that introduces an inverse Wishart prior on $\boldsymbol{\eta}$ using the current WISDM-68 data, we do in fact observe a covariance structure very similar to what is observed when to the covariance matrix is calculated from the $\boldsymbol{\eta}$ estimates with independent univariate priors.

negative end). Importantly, the specific forms of response style observed among individual respondents may not align with only one component, but rather a combination of components, as the components are mutually orthogonal. Thus, while there are three distinguishable dimensions that appear useful in describing response style heterogeneity, the components have the potential to work in combination to reflect many different response style types.

Recall that our analysis also chose to treat missing responses as ‘4’s. While at one level this treatment seemed appropriate (to the extent that missing might correspond to uncertainty, and hence an anticipated response in the middle of the rating scale), this naturally leads to a confounding of MRS with nonresponse. We examined the implications of our treatment of missing responses as ‘4’s by contrasting the principal component analysis results with an analysis that used listwise deletion in treatment of missing responses (reducing the sample size to 1437). We observed virtually identical results, suggesting no noticeable effects on our overall evaluation of response-style heterogeneity. Nevertheless, we note that in practice alternative approaches might be taken. For example, one possibility would be to treat the missing response as a unique response category, and empirically estimate its relationship to the substantive trait.

Table 4 displays the resulting component scores for the respondents in Table 2. As the component scores are standardized, Misfit respondents 1472, 1228 and 720 provide examples of response styles that appear to be captured to some extent by these three components given the presence of component scores that are large in absolute magnitude; misfit respondents such as 230 and 1431 are somewhat less so, likely reflecting the uniqueness of their response styles. Respondent 1228 provides an example in which all three components appear relevant, with the three component scores all being of at least moderate size in absolute magnitude. Recall that while Respondent 1228 shows evidence of anti-ERS (negative on Component 1), the Respondent also shows a tendency to under-select 2 (negative on Component 2) and 6 (negative on Component 3). In other words, the presence of three meaningful components makes possible unique response style types that are best reflected by some combination of the components. In summary, the presence of three meaningful (but still incomplete) principal components gives a clear sense that there are many different response style types present in the WISDM-68 data.

Taken together, response-style heterogeneity appears quite common with the category frequency person fit statistic indicating that the majority of respondents showing detectable departures in rating scale usage from expectations under a partial credit item response model. Further, these results suggest that while response styles such as ERS emerge as sources of response-style heterogeneity, other response styles also exist. If some response styles (e.g., unexpected ones) are not identified and addressed, they might confound the measurement of identified response styles and the substantive trait(s). Given these results, we look next at some of the ways in which the model in (1) can provide a basis for better understanding the implications of response-style heterogeneity on measurement precision.

Study 2: Evaluating the Implications of Response-Style Heterogeneity for Reduced Measurement Precision

By assuming a simultaneous influence of substantive and response-style trait(s) on category selection, the MNRM approach to response-style heterogeneity also provides a tool for evaluating the effects of response style on the precision of measurement related to the intended substantive trait(s).

Study 2 Goals

We have previously addressed the use of the MNRM to study bias (Bolt & Johnson, 2009; Bolt, Lu, & Kim, 2014) so we focused primarily on implications for precision in this study. We anticipated that varying levels of heterogeneity in respondent response styles would lead to varying effects on the estimated precision of a respondent's trait estimates, as the more respondents vary in their use of rating scales, the less confident one becomes that the substantive trait estimate is a meaningful reflection of the underlying trait. Given the adoption of a Bayesian estimation approach, we evaluated the precision of substantive and response-style trait estimates by attending to the posterior standard deviations (psd's) for each respondent parameter.

Study 2 Method

We initially examined the effects of response-style heterogeneity through simulation analyses. Specifically, using the category intercept estimates from the WISDM-68 analysis as generating item parameters, we simulated respondent trait parameters for 1000 respondents according to the priors specified earlier (i.e, $\theta_j \sim Normal(0,1)$), but consider separate conditions of (1) low: $\eta_{jk} \sim Normal(0,.08)$, (2) medium: $\eta_{jk} \sim Normal(0,.75)$ and (3) high $\eta_{jk} \sim Normal(0,2.08)$ response-style heterogeneity. We fitted the MNRM using the same Bayesian specifications as applied above. Our initial goal was to examine how well the model recovered σ^2_η and distinguished conditions of low, medium, and high response-style heterogeneity.

Study 2 Results

We found the MCMC algorithm recovered σ^2_η quite well, with $\hat{\sigma}^2_\eta = .09$ (se=.003) in the low heterogeneity condition, $\hat{\sigma}^2_\eta = .76$ (se= .011) in the medium heterogeneity condition, and $\hat{\sigma}^2_\eta = 2.15$ (se= .037) in the high heterogeneity condition. We next examined the implications of these varying levels of response-style heterogeneity on the respondent parameter estimates. Specifically we examined the posterior standard deviations of both the θ and η parameters. Not surprisingly, we found the posterior standard deviations of both the response-style and substantive trait parameters to become increasingly larger under the conditions of medium and high response-style heterogeneity. As anticipated, greater response-style heterogeneity coincided with greater uncertainty both in the estimated response styles of respondents, and as a consequence, in the estimated substantive trait.

Figure 1 illustrates the relationship between $\hat{\theta}$ and psd(θ) under the low, medium, and high response-style heterogeneity conditions. As implied by Table 5, psd's increased as heterogeneity increased. In a Bayesian estimation framework, the greater uncertainty in $\hat{\theta}$ also yielded shrunken estimates, as greater response-style heterogeneity resulted in estimates that are also pulled closer to the mean (0). This is seen in Figure 1 where the extreme $\hat{\theta}$s (in each condition, $\theta_j \sim Normal(0,1)$) were successively pulled closer to 0 as the amount of response-style heterogeneity increases. Also of interest is the increased variability seen in psd(θ) for the medium and high heterogeneity conditions conditional upon $\hat{\theta}$.

In a practical measurement setting, one application of these results would be to quantify the anticipated value of adopting an anchoring vignette design, for example. A goal of administering anchoring vignettes is to estimate examinee response style ($\hat{\eta}$) independent of the self-report items, and to, in turn, use those estimates to improve estimation of $\hat{\theta}$ (Bolt, Lu, & Kim, 2014). In this regard, we can also evaluate the posterior standard deviations of $\hat{\theta}$ under a condition where each η is fixed. The last row of Table 5 shows this estimate when each η was set at its known value of 0. The resulting value functions as a type of limiting condition against which to evaluate the effects of the response-style heterogeneity.

Consistent with the relation between response-style heterogeneity and the precision of the estimated substantive trait, we also saw a strong relationship between the precisions of response-style estimates and substantive trait estimates at the individual respondent level. Figure 2 displays the relationship observed for each of the simulation conditions and the real data. The magnitude of response style (as quantified by the average absolute magnitude of $\hat{\eta}$ across categories), as anticipated, contributed to a reduction in the precision of the estimated θ. As an illustration of this, the last two cases in Table 2 (Respondents 180, 1431) show example response patterns with similar overall observed θ estimates under the PCM model (−.40 and −.50 respectively), both having posterior standard deviations of .08. But under the MNRM, the posterior standard deviation of Respondent 1431 increased much more (.31) than that of Respondent 180 (.18). Consistent with the explanation above, the increase in posterior standard deviation observed for Respondent 1431 was possibly largely caused by the presence of a strong response style, in this case a tendency to over-select Category 3.

Study 3: Using a MIRT Model of Response Style to Inform Test Design

A third aspect of response-style influence that we can examine using an MNRM relates to the effects of instrument design. Many self-report instruments consist of items that are approximately parallel in the sense of the classical true score model(i.e., items with the same true scores, variances for each respondent). While item parallelism is sometimes viewed as a desirable psychometric property, in the presence of response styles, item parallelism might be anticipated as problematic.

Study 3 Goals

In this section we use the MNRM to explore how test design interacts with response style to affect (1) the bias of sum scores as indicators of trait level, and (2) measurement precision.

We anticipated that greater item parallelism yields more substantial effects of response style, and use the MNRM model of response style to better understand these effects.

Study 3 Method

Using a simulation illustration, we generated response data in relation to two hypothetical tests. Specifically, we considered two hypothetical tests that are distinguished by the amount of variability simulated in the $\boldsymbol{c}$ vectors across items. Under an MNRM with equal category slopes, we can created a condition of item parallelism by holding constant the $\boldsymbol{c}$ parameter vectors across items, while a condition of item heterogeneity (i.e., lack of parallelism) exists to the extent that the $\boldsymbol{c}$ parameter vectors vary. For the item parallelism condition, we generated $c_k \sim Normal(0,1)$ using the same category intercepts across items, while for the item heterogeneity condition, we generated $c_{ik} \sim Normal(0,1)$, allowing the category parameters to vary across items.

For simplicity, we generate data with respect to a special case of the MNRM in which response-style heterogeneity is only a function of an ERS dimension. As above, the respondent parameters were generated to vary with respect to both a substantive trait dimension (θ) and a response-style dimension corresponding to ERS (η_{ERS}). The model used to generate the data is from Bolt and Johnson (2009):

$$P(X_{ij} = \mathrm{k} \mid \theta_j; \eta_j) = \frac{\exp(a_{ik}\theta_j + c_{ik} + a_{ERSi,k}\eta_{ERS,j})}{\Sigma_{g=1}^{l} \exp(a_{ig}\theta_j + c_{ig} + a_{ERS,i,g}\eta_{ERS,j})} \quad (3)$$

where $\boldsymbol{a}_i = (-3,-2,-1,0,1,2,3)$ and $\boldsymbol{a}_{ERS,i} = (2,-.8,-.8,-.8,-.8,-.8,2)$for all items i, again applying normalization constraints $\Sigma_{g=1}^{l} c_{ig} = 0$ within items. Further, we generated $\theta_j \sim Normal(0,1)$, and $\eta_{ERS,j} \sim Normal(0,1)$. Item responses were simulated for 50 7-category items and 1000 respondents. From the previous section, the capability of measuring θ should naturally be dependent on how well η_{ERS} is measured. The ability to measure η_{ERS} is likely also affected by test design. We expected, for example, that item parallelism will increase the difficulty in determining whether the consistent use of the same response category reflects a response-style tendency or not, as item parallelism would imply that the same item scores tend to be observed across items even in the absence of response style. We thus also examined the precision with which η_{ERS} was estimated for each test under the MNRM. Along the lines of the previous section, we also compared the two hypothetical tests with respect to their estimation of θ.

Study 3 Results

Table 6 displays the results. As anticipated, under the conditions of item parallelism, we observed substantially less precision in the estimation of η_{ERS}, and consequently, a much greater increase over the PCM in terms of the mean posterior standard deviation of θ. Further inspection of the distribution of these mean posterior standard deviations across respondents suggested the difference can be attributed in part to several patterns for which posterior standard deviations are very large in the parallel item condition.

Following Bolt and Johnson (2009), we also estimated bias in the expected sum score for the scale by evaluating the expected sum score as a function of θ, η_{ERS} and compared it against the expected sum score when η_{ERS} is at its mean (0). The expected sum score was calculated as a function of the category probability functions defined by the model in (3) where

$$ES(\theta, \eta_{ERS}) = \Sigma_{i=1}^{NI} \Sigma_{k=1}^{l} k \times P(U_i = k \mid \theta, \eta_{ERS}) \quad (4)$$

and then the difference in expected scores was calculated as:

$$BIAS(\theta, \eta_{ERS}) = ES(\theta, \eta_{ERS}) - ES(\theta, 0) \quad (5)$$

to quantify bias. Figure 3 illustrates the resulting bias curves across different levels of η_{ERS} for the measures with item parallelism and item heterogeneity, respectively. As seen in the figure, greater heterogeneity of items tended to reduce the concentration of bias related to ERS. For the specific parallel test simulated, this bias is found to be maximized just below and just above the mean θ level (0). At such θ locations, expected scores of 3s (for respondents having θ just below 0) or 5s (for respondents having θ just above 0) are anticipated (on a 7 point rating scale); however, high ERS respondents would instead select 1s and 7s, respectively, thus creating the substantial bias in overall sum scores.

Thus, item heterogeneity appears to function in two positive ways with regard to response styles on self-report survey instruments. First, it makes response styles easier to detect/measure, and thus potentially control for, in terms of their influence on the substantive trait. Second, heterogeneous tests appear to reduce bias associated with ERS, or at a minimum, reduce its concentration. In the above example, where the most substantial bias occurs near the center of the θ distribution, it seems that this reduced concentration of bias is particularly beneficial.

Implications for Response-Style Assessment for the WISDM-68

Based on the analyses conducted in this paper, we arrived at several conclusions regarding response-style influence on the WISDM-68. First, response style appeared to come in a variety of forms. Variability in the use of extreme and midpoint response categories explained part of this heterogeneity, although other sources of heterogeneity are present as well. The overall amount of response-style heterogeneity was of medium size ($\hat{\sigma}_{\eta}^2 = 0.745$) relative to the conditions simulated; further, the effect on MNRM posterior standard deviations of the substantive trait was of reasonable magnitude (increasing on average from .08 under the PCM to .23 under the MNRM), suggesting that accounting for response-style heterogeneity influenced the precision with which the substantive trait was estimated.

Second, aspects of test design appear to influence the effects of response style. An appealing feature of the WISDM-68 is that its items lack parallelism with items that vary considerably in terms of their means (2.67 to 5.92), implying that most respondents are expected to report scores reflecting a range of different rating-scale values. As a result, the effects of response style in terms of bias were reduced, and likely confined to respondents showing strong response styles. Figure 4 shows examples of bias curves related to two of the respondents in

Table 2—Respondent 1472, who showed evidence of ERS, and Respondent 1431, who over-selected 3s. Of all 1504 respondents, the largest adjustment in θ comparing the PCM and MNRM was seen for Respondent 1431, whose θ estimate was still only adjusted (decreased) by ½ standard deviation (from −.50 to −.97), a relatively small adjustment. Nevertheless, the reason for the adjustment is apparent from the bias curve for Respondent 1431, which achieved its peak at approximately $\theta = -1.3$, close to the MNRM θ estimate of the respondent (and, as the adjustment would suggest, yields bias in a positive direction). The bias curve for the response style of Respondent 1472 looked similar to that implied by a heterogeneous test (as in the simulation). Apart from a few individual respondent adjustments, the adjusted MNRM θ estimates still correlated with the PCM θ estimates at a high level of .95, and with WISDM-68 total scores at .92, suggesting the overall biasing effects were relatively small.

A third feature of the WISDM-68 dataset with regard to response style is that we observed only minimal correlations between the substantive trait and response styles. Specifically, the correlations between the MNRM θ and each of Components 1, 2 and 3 were .07 ($p = .008$), .00 ($p = .862$), and .05 ($p = .075$), respectively, and range from −.13 ($p < .001$) for η_3 to .12 ($p < .001$) for η_7. The weakness of such correlations generally implies minimal effects in terms of correlational bias from use of the WISDM-68 (Plieninger, 2016). Indeed, we found that use of the bias-corrected MNRM θ estimates has virtually no effect on the prediction of abstinence at either end-of-treatment or six months post-quit. Of course, while these abstinence outcomes are highly clinically important, they are likely insensitive to dependence per se given that they are influenced by numerous additional factors (Bolt et al., 2009).

Despite these encouraging findings for WISDM-68, we note that there are contexts where the effects of response style on the precision of substantive trait estimates can have implications. Bolt and Johnson (2009) and Morren, Gelissen, and Vermunt (2012), for example, examined how response-style heterogeneity can frequently contribute to apparent differential item functioning (DIF), particularly when response styles correlate with other respondent variables. In this respect, all three PCs of response style show correlation with education (−.26, $p < .001$; .20, $p < .001$; .18, $p < .001$, respectively), suggesting potential interference of response styles in DIF analyses related to education.

Conclusion

A goal of this paper was to highlight the potential of an MNRM model of response style to inform how response-style heterogeneity may interfere with measurement of a substantive trait. The flexibility afforded by an MNRM in representing many different response styles makes it potentially useful both in examining response styles at the respondent level as well as in quantifying the overall amount of response-style heterogeneity in a test-taking population. As a statistical model, the MNRM can be used to understand not only how response styles contribute to bias, but also how they may reduce the precision with which the intended substantive trait is measured.

When using a fully Bayesian method to estimate the model, we observed that the nature of the interference produced by response styles depended on individual respondent-, respondent population-, and test-related factors. Specifically, individual respondents whose score category usage reflects a greater departure from the norm (as reflected by a larger absolute η) tended to show greater increases in the posterior standard deviation of θ. Likewise, greater response-style heterogeneity in the population, as quantified by the response-style hyperparameter variance $\hat{\sigma}^2_{\eta}$, similarly led to greater uncertainty, both in individual response styles, and as a consequence, in the substantive trait. Finally, in terms of test properties, greater item parallelism yielded similar reductions in response-style and substantive trait precision. Tests with greater item heterogeneity seemed to provide an advantage both in allowing for better assessment of response style (and in turn the substantive trait), but also appeared to simultaneously disperse the concentration of bias that can occur as a result of response styles such as ERS. For the WISDM-68, the biasing effects of response style seemed to be mitigated both by the presence of item heterogeneity as well as the lack of meaningful correlations between the substantive trait and more prominent response-style types (e.g., ERS).

While the MNRM is flexible, it is also important to acknowledge its limitations, especially its inability to detect response styles that correlate highly with θ. This limitation may be significant in that such response styles ultimately seem to contribute the most to bias (Plieninger, 2016). As an example, acquiescent tendencies may be more difficult to detect with the MNRM, especially for tests where item endorsement (agreement) is also simultaneously indicative of higher θ. Additional design considerations beyond item heterogeneity, such as reverse worded/scored items, may be needed to help make such response styles more apparent.

At a minimum, it seems that response-style heterogeneity of medium or greater magnitude should call into question the application of standard IRT models. From one perspective, response styles can be viewed as a source of local dependence that implies individual items should not be viewed as independent sources of information about the trait. The analyses presented in this paper were designed to be illustrative; clearly more extensive real data and simulation studies will be useful in both extending the findings, and addressing some of the inevitable nuances associated with this work. It will be useful to consider generalizations of the model that permit correlations between both the substantive and response-style latent traits, both to maximize interpretability of the response styles as well as to yield a substantive trait that is maximally aligned with the intended-to-be-measured trait.

References

Austin EJ, Deary IJ, & Egan V (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. Personality and Individual Differences, 40, 1235–1245. doi:10.1016/j.paid.2005.10.018

Baumgartner JEM, & Steenkamp H (2001). Response style in marketing research: A cross-national investigation. Journal of Marketing Research, 38, 143–156.

Böckenholt U (2012). Modeling multiple response processes in judgment and choice. Psychological Methods, 17, 665–678. doi:10.1037/a0028111 [PubMed: 22545594]

Böckenholt U, & Meiser T (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. British Journal of Mathematical and Statistical Psychology, 70(1), 159–181. [PubMed: 28130934]

Bolt DM, Piper ME, McCarthy DE, Japuntich SJ, Fiore MC, Smith SS, & Baker TB (2009). The Wisconsin Predicting Patients' Relapse questionnaire. *Nicotine &* Tobacco Research, 11(5), 481–492. PMCID: PMC2671459.

Bolt DM, & Johnson TR (2009). Applications of a MIRT model to self-report measures: Addressing score bias and DIF due to individual differences in response style. Applied Psychological Measurement, 33, 335–352.

Bolt DM, & Newton JR (2011). Multiscale measurement of extreme response style. Educational and Psychological Measurement, 71, 814–833.

Bolt DM, Lu Y, & Kim JS (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. Psychological Methods, 19(4), 528–541. [PubMed: 24773355]

DeBoeck P, & Partchev I (2012). IRTrees: Tree-based item response models of the GLMM family. Journal of Statistical Software, 48(1), 1–28. doi:10.18637/jss.v048.c01

Eid M, & Rauber M (2000). Detecting measurement invariance in organizational surveys. European Journal of Psychological Assessment, 16, 20–30. doi:10.1027//1015-5759.16.1.20

Falk CF, & Cai L (2016). A flexible full-information approach to the modeling of response styles. Psychological Methods, 21, 328–347. [PubMed: 26641273]

Gelman A & Rubin DB (1992). Inference from iterative simulation using multiple sequences. Statistical Science, 11, 457–472. 10.1214/ss/1177011136

Glas CAW, & Meijer RR (2003). A Bayesian approach to person fit analysis in item response theory, Applied Psychological Measurement, 27, 217–233.

Greenleaf EA (1992). Improving rating scale measures by detecting and correcting bias components in some response styles. Journal of Marketing Research, 29, 176–188.

Jeon M & De Boeck P (2016). A generalized item response tree model for psychological assessments. Behavior Research Methods, 48, 1070–1085. [PubMed: 26208813]

Khorramdel L, & von Davier M (2014). Measuring response styles across the big five: A multiscale extension of an approach using multinomial processing trees. Multivariate Behavioral Research, 49, 161–177. doi:10.1080/00273171.2013.866536 [PubMed: 26741175]

Kieruj ND, & Moors G (2012). Response style behavior: Question format dependent or personal Style? Quality & Quantity . doi: 10.1007/s11135-011-9511-4.

King G, Murray CJL, Salomon JA, & Tandon A (2004). Enhancing the validity and cross-cultural comparability of survey research. American Political Science Review, 98, 191–207. doi:10.1017/S000305540400108X.

Masters GN (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149–174.

Morren M, Gelissen J, & Vermunt J (2012). The impact of controlling for extreme responding on measurement equivalence in cross-cultural research. Methodology, 8, 159–170. doi: 10.1027/1614-2241/a000048.

Piper ME, Felderman EB, Piasecki TM, Bolt DM, Smith SS, Fiore MC, & Baker TB (2004). A multiple motives approach to tobacco dependence: The Wisconsin Inventory of Smoking Dependence Motives (WISDM). Journal of Consulting and Clinical Psychology, 72, 139–154. [PubMed: 15065950]

Plieninger H (2016). Mountain or molehill? A simulation study on the impact of response styles. Educational and Psychological Measurement, 77, 32–53. doi: 10.1177/0013164416636655 [PubMed: 29795902]

Plieninger H, & Meiser T (2014). Validity of multiprocess IRT models for separating content and response styles. Educational and Psychological Measurement, 20, 1–25.

Sinharay S (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach, Journal of Educational Measurement, 42, 375–394.

Spiegelhalter D, Thomas A, & Best N (2003). WinBUGS version 1.4 user manual. Cambridge, England: MRC Biostatistics Unit.

Thissen-Roe A, & Thissen D (2013). A two-decision model for responses to Likert type items. Journal of Educational and Behavioral Statistics, 38, 522–547. doi:10.3102/1076998613481500

Van Rosmalen J, Van Herk H, & Groenen PJF (2010). Identifying response styles: A latent-class bilinear multinomial logit model. Journal of Marketing Research, 47, 157–172.

Van Vaerenbergh Y & Thomas TD (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. International Journal of Public Opinion Research, 25, 195–217.

von Davier M, & Rost J (1995). Polytomous mixed Rasch models In Fischer GH & Molenaar IW (Eds.), Rasch models – Foundations, recent developments and applications (pp. 371–379). New York, NY: Springer.

von Davier M, Shin H-J, Khorramdel L, & Stankov L. (2017). The effects of vignette scoring on reliability and validity of self-reports. Applied Psychological Measurement, 42, 291–306. [PubMed: 29881126]

Wetzel E, & Carstensen CH (2015). Multidimensional modeling of traits and response styles. European Journal of Psychological Assessment. Advance online publication. doi: 10.1027/1015-5759/a000291.

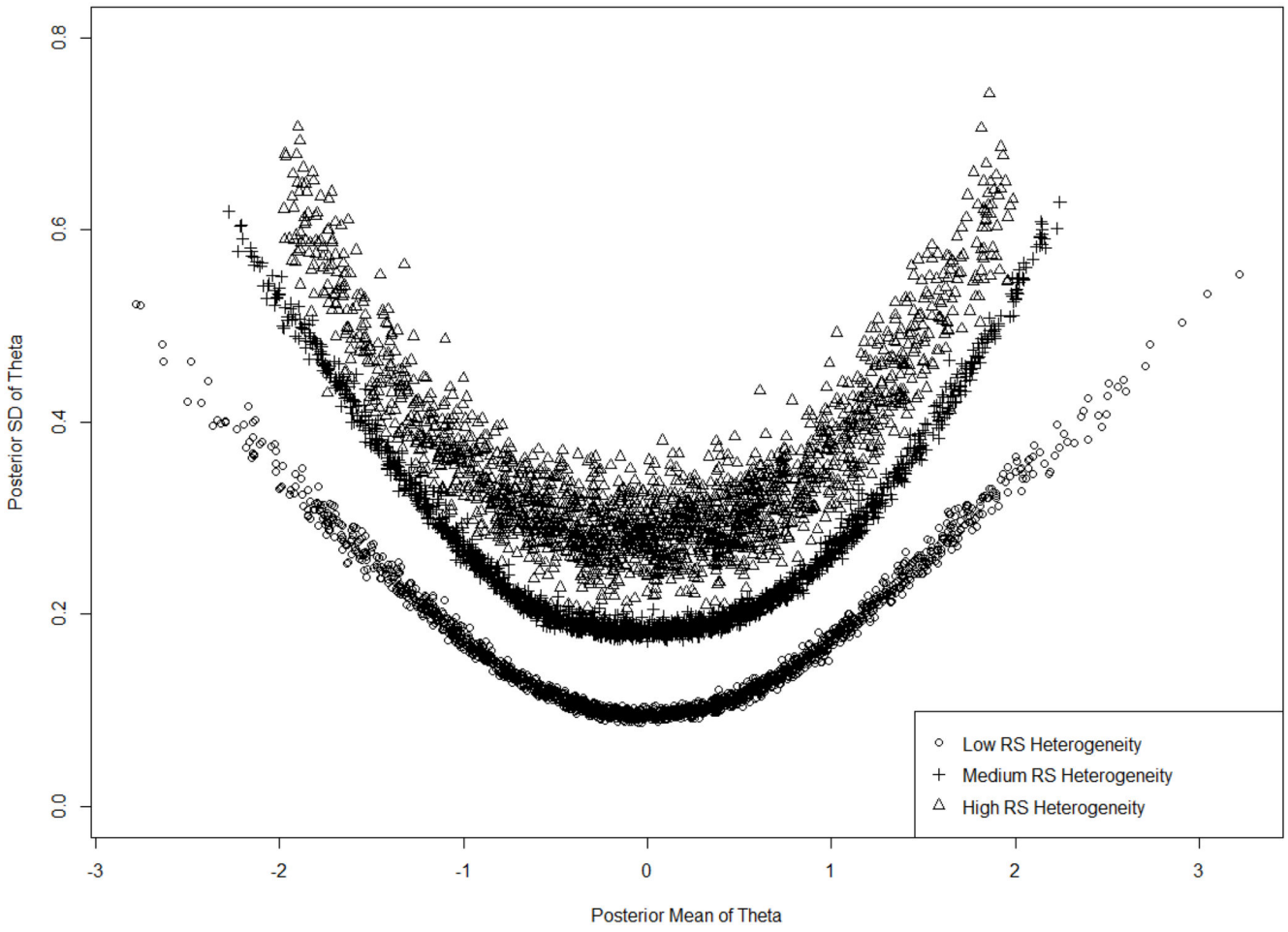

Figure 1.
Scatterplot of Posterior Standard Deviation against Posterior Mean of Theta Estimates, Low, Medium, and High Response-Style Heterogeneity Conditions, Simulation Study.

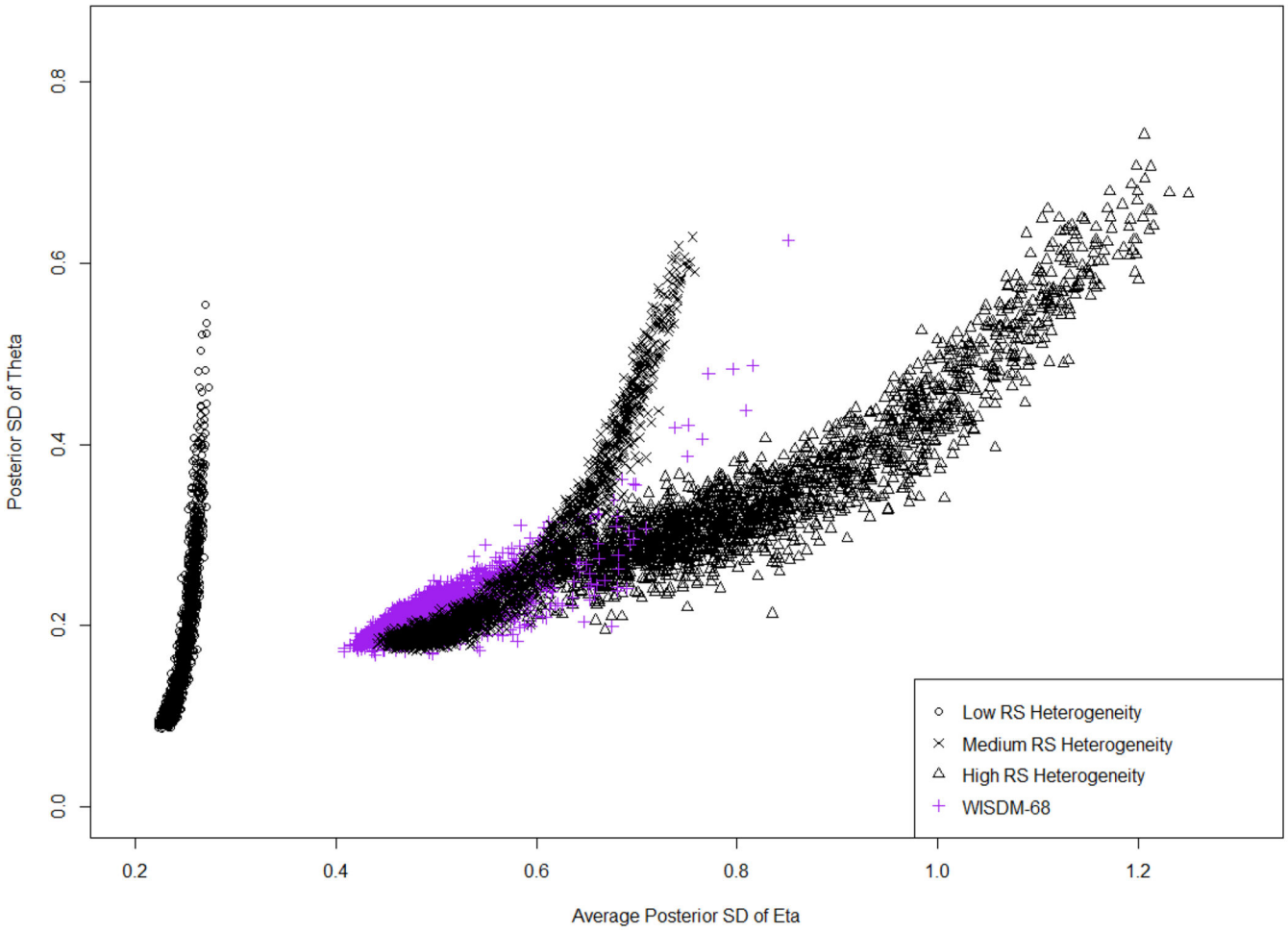

Figure 2.

Scatterplot of Posterior Standard Deviation of Theta Estimates Against Average Posterior Standard Deviation of Eta Estimates, Low, Medium, and High Response-Style Heterogeneity Conditions, Simulation Study and Real Data Study.

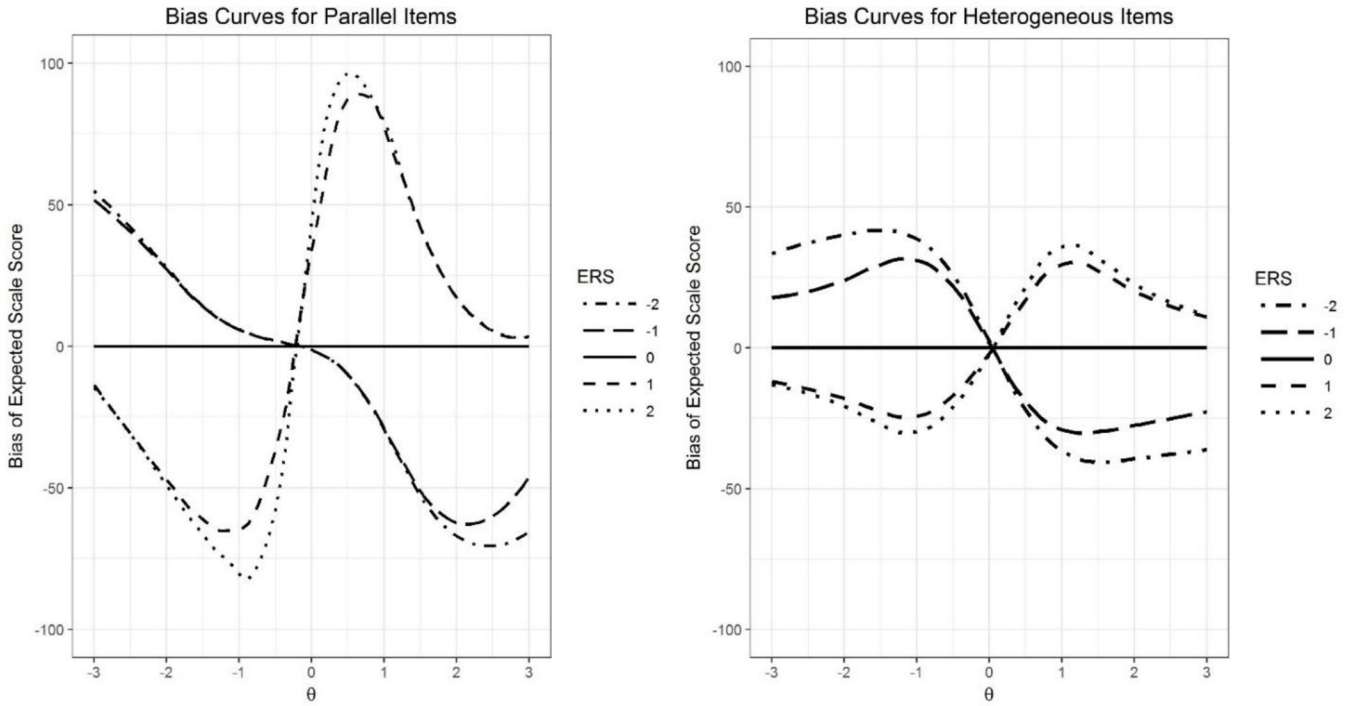

Figure 3.

Quantification of Bias in Expected Scale Score in Relation to Theta as a Function of Level of ERS, Hypothetical Parallel Item and Heterogeneous Item Tests.

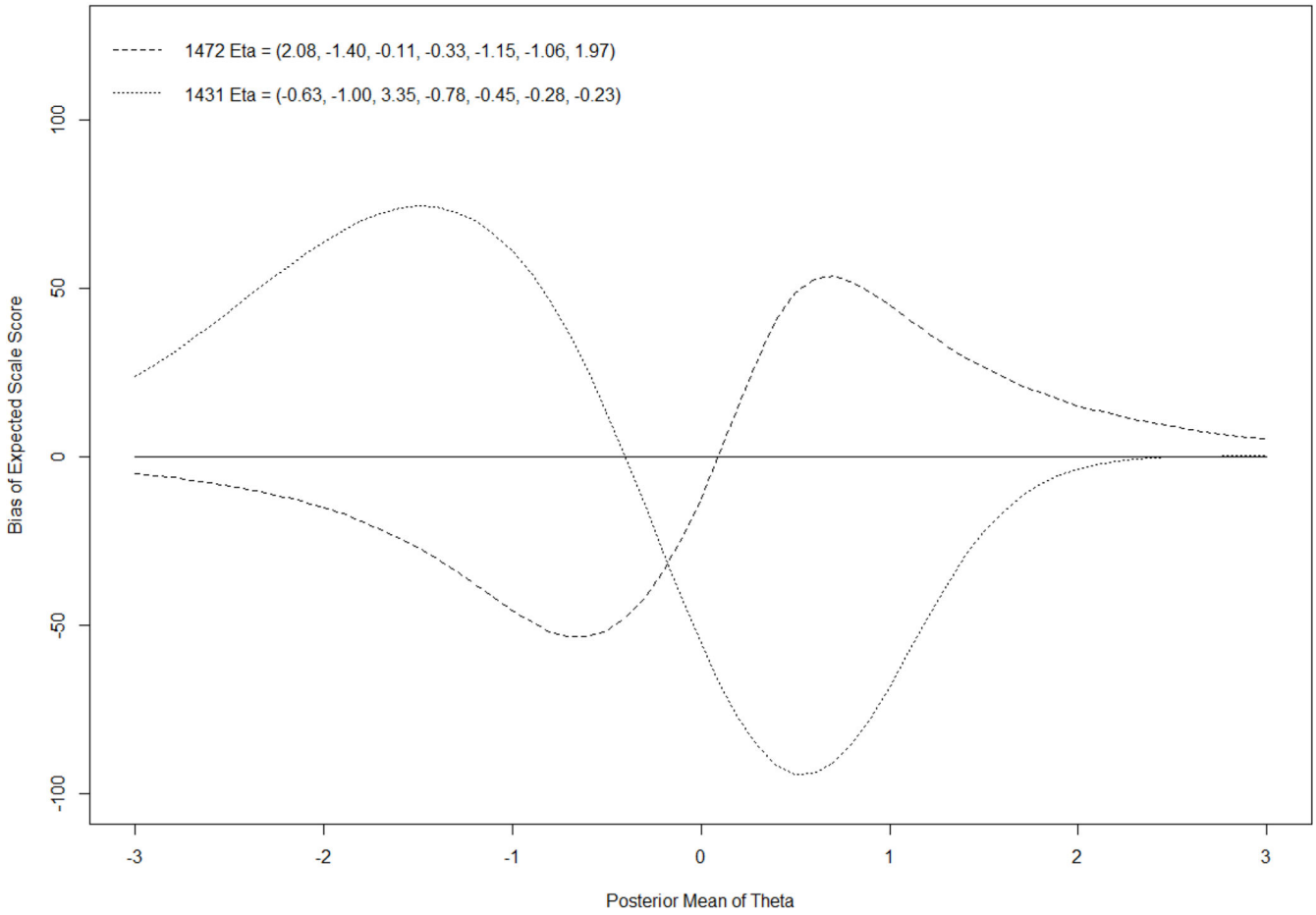

Figure 4.
Quantification of Bias in Relation to Theta as a Function of Observed Eta Estimates, WISDM-68 test, Respondents 1431 and 1472.

Table 1

Cross-Tabulation of Person Misfit Frequencies with respect to the T_L and T_F Deviance Statistics, WISDM-68 respondents (N=1504)

	Category Frequency (T_F)			
Likelihood (T_L)	**Under-fit**	**Fit**	**Over-fit**	**Row marginal**
Under-fit	92	47	1	140
Fit	710	589	2	1301
Over-fit	19	44	0	63
Column marginal	821	680	3	1504

Table 2.

Example Response Patterns, their Corresponding T_L and T_F p-values, and Estimated MNRM Parameters, 1504 WISDM-68 respondents

		p-value (PCM)		MNRM							
Respondent	**Item Response Pattern**	T_L	T_F	$\hat{\theta}$ (se)	$\hat{\eta}_1$ (se)	$\hat{\eta}_2$ (se)	$\hat{\eta}_3$ (se)	$\hat{\eta}_4$ (se)	$\hat{\eta}_5$ (se)	$\hat{\eta}_6$ (se)	$\hat{\eta}_7$ (se)
1472	711373777774314111 117111714173171111 111171111111114117 11111111111111	0.0949	0.0000 *	−0.67 (0.09)	2.08 (0.66)	−1.40 (0.65)	−0.11 (0.48)	−0.33 (0.47)	−1.15 (0.64)	−1.06 (0.69)	1.97 (0.71)
1228	443644364444444343 434434454345444443 434353354344344545 54444445534433	0.0793	0.0000 *	−0.06 (0.07)	−1.23 (0.74)	−1.26 (0.68)	1.54 (0.40)	2.22 (0.26)	0.73 (0.41)	−0.59 (0.61)	−1.41 (0.74)
720	632766277276272126 642242666662276664 662666426274146227 62276616766442	0.9684	0.0000 *	0.15 (0.07)	−0.32 (0.61)	1.35 (0.46)	−1.07 (0.55)	0.08 (0.34)	−1.62 (0.60)	1.25 (0.44)	0.33 (0.60)
230	374674357474464723 742144775771744626 633477372223277227 67322277737422	0.8255	0.0023 *	0.15 (0.07)	−0.61 (0.65)	0.99 (0.47)	0.34 (0.37)	0.34 (0.27)	−1.20 (0.46)	−0.57 (0.48)	0.70 (0.60)
470	712377666677611275 573774666777177774 644351777475363777 74143746534656	0.0040 *	0.1630	0.38 (0.08)	0.59 (0.70)	−0.56 (0.60)	0.19 (0.40)	−0.05 (0.31)	−0.57 (0.38)	−0.09 (0.46)	0.50 (0.66)
582	611552221163121111 1117312141471263 31 11414172111126632 11115214213211	0.3187	0.2940	−0.68 (0.09)	0.54 (0.61)	−0.06 (0.45)	−0.37 (0.40)	−0.37 (0.40)	−0.39 (0.52)	0.36 (0.56)	0.29 (0.68)
180	242315346441432234 134143311471532741 424516321213144126 1442224625122	0.1974	0.2760	−0.46 (0.18)	−0.26 (0.50)	0.12 (0.39)	0.01 (0.32)	0.65 (0.26)	−0.34 (0.42)	0.11 (0.49)	−0.30 (0.62)
1431	3333311121333333333 333333333133133333 333333333333333333 33333333333333	0.0024 *	0.0000 *	−0.97 (0.31)	−0.63 (0.73)	−1.00 (0.59)	3.35 (0.42)	−0.78 (0.71)	−0.45 (0.81)	−0.28 (0.86)	−0.23 (0.84)

Notes:

* p<.05;

PCM = Partial Credit Model; MNRM = Multidimensional Nominal Response Model

Table 3

Component Loadings for First Three Components, Principal Components Analysis of η Estimates at Respondent Level

	Rescaled Component Loadings		
Category	**Component 1**	**Component 2**	**Component 3**
1	.93	−.26	.24
2	−.42	.85	.02
3	−.61	−.03	−.50
4	−.50	−.44	−.43
5	−.60	−.45	.20
6	−.41	.14	.80
7	.93	.21	−.29

Table 4

Component Scores of Principal Components 1-3 for the Example Respondents in Table 3

	Component Scores		
Category	**Component 1**	**Component 2**	**Component 3**
1472	2.57	−.96	−1.17
1228	−1.74	−2.89	−2.29
720	.25	2.76	1.13
230	.15	1.93	−1.75
470	.73	−.27	−.33
582	.61	.34	.72
180	−.27	.05	−.28
1431	−.64	−.27	−2.18

Table 5

Average Estimated Posterior Standard Deviations of MNRM Respondent Parameters, Low, Medium and High Response-Style Heterogeneity Conditions, Simulation Analyses

	Mean Posterior Standard Deviation	
	η	θ
Low Heterogeneity	0.24	0.16
Medium Heterogeneity	0.57	0.26
High Heterogeneity	0.83	0.36
η Fixed		0.14

Table 6

Descriptive Statistics for Posterior Standard Deviations of θ, Item Parallelism and Item Heterogeneity Conditions

	Mean Posterior Standard Deviation			SD of Posterior Standard Deviations (Min,Max)		
	PCM	MNRM		PCM	MNRM	
Condition	$\hat{\theta}$	$\hat{\theta}$	$\hat{\eta}_{ERS}$	$\hat{\theta}$	$\hat{\theta}$	$\hat{\eta}_{ERS}$
Parallel	.17	.25	.36	.08(.11,.46)	.13(.06,.72)	.19(.08,.65)
Heterogeneous	.14	.18	.21	.08(.09,.51)	.09(.09,.55)	.11(.10,.54)